**Project1 Report**

**Task 4)**

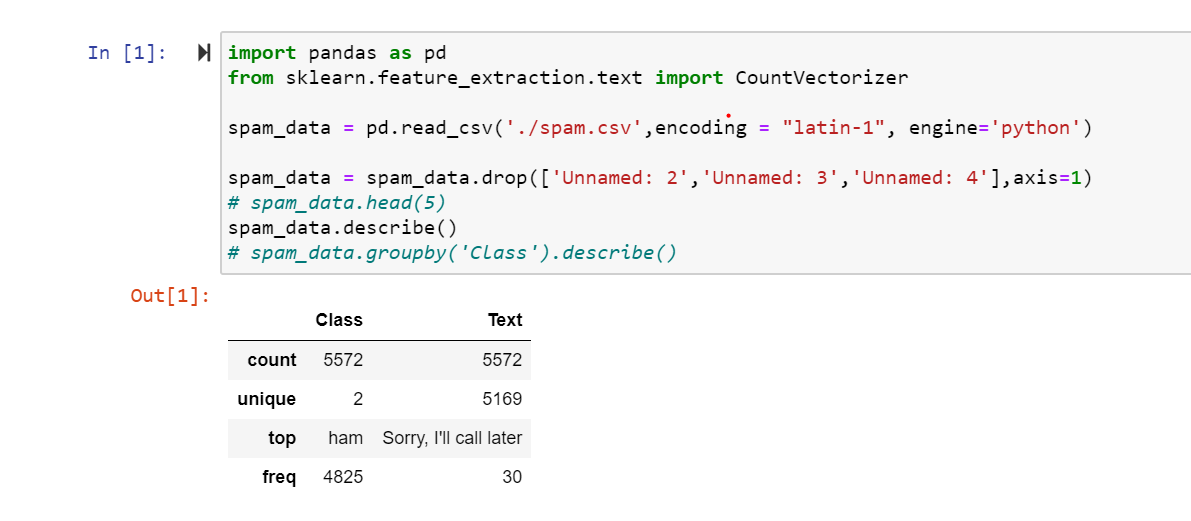
The objective of fourth task is to clean the given dataset and evaluate the model by applying the different techniques TFIDF, Count\_Vectorizer and then evaluate the best algorithm by analyzing the results

**Code Explanation :**

Imported the required libraries

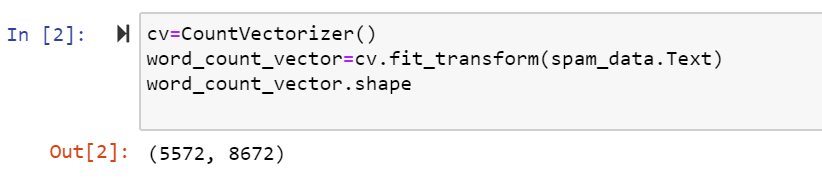
Imported the given spam.csv and encoded it and then dropped the unnecessary columns

And then described the data as shown in the below screen capture



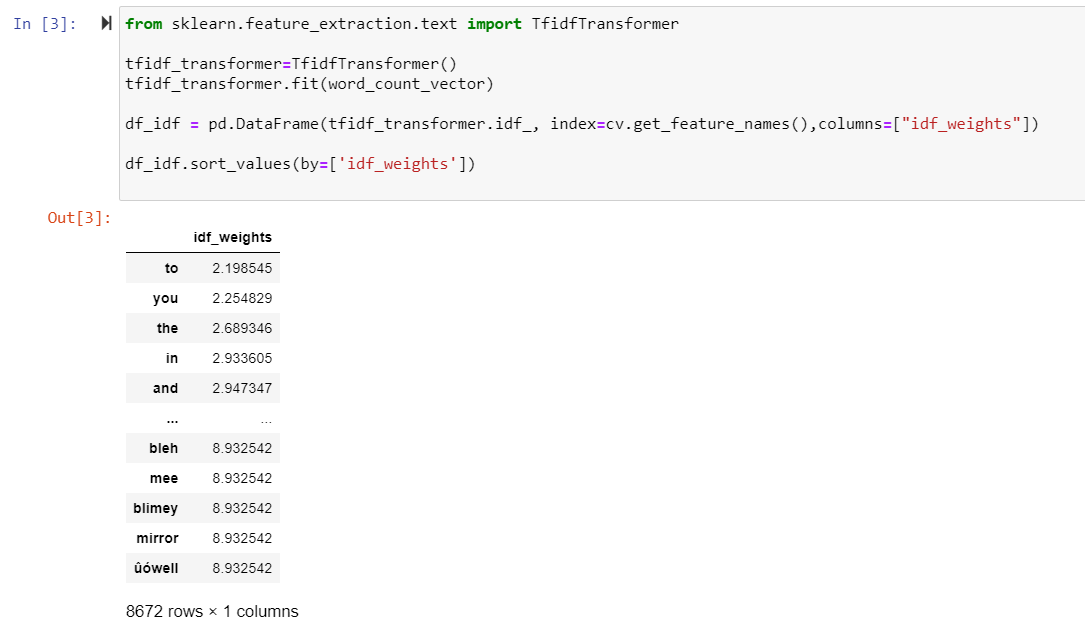
Initialize the count vectorizer and fit the data

And then get into a variable word\_count\_vector



Using TfidTransformer algorithm got the most frequently used words in the document

Idf\_weights gives the least value for most frequently used word



Transform spam\_data.Text into count\_vector

Declared a variable ‘h’ to loop through the document to get the frequency of most important words

Sort the values by tfidf

Restricted loop to 3 iterations as there are lot of rows in the document



Finally trained and tested the data

Split the data by Text and Class

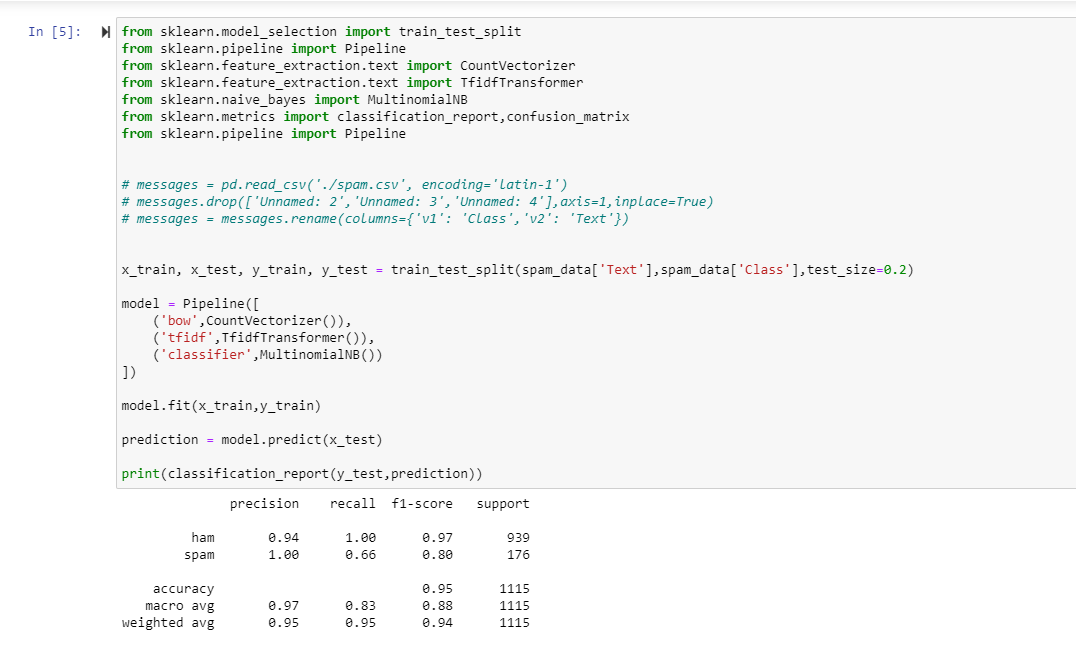
Applied three algorithms CountVectorizer, TfidTransformer, MultinomialNB

Fit the data into model

Predict the data by passing x\_test

Print the classification of the document

Printed the precision values of class and Text



After analyzing the results CountVector gives the better results compared to TFIDF vectorizer

MultinomialNB performed best.

**Panopto Recording:**

<https://umkc.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=e7ce6baf-7317-44a3-8f4a-abf1004a239a>